Modeling Trends of Airborne Particulate Matter by using Support Vector Machines

Artemio Sotomayor-Olmedo, M. Antonio Aceves-Fernandez, Efren Gorrostieta-Hurtado, J. Carlos Pedraza-Ortega, J. Manuel Ramos-Arreguin, J. Emilio Vargas-Soto, Saul Tovar-Arriaga.
Facultad de Informatica, Universidad Autonoma de Queretaro,
Queretaro, Mexico.
artemiosotomayor@gmail.com,marco.aceves@gmail.com,efrengorrostieta@gmail.com,caryoko@yahoo.com,jramos@mecamex.net,emilio@mecamex.net,saulotov@yahoo.com.mx

Abstract— **Monitoring, modeling and forecasting of air quality parameters are important topics in environmental and health research due to their impact caused by exposing to air pollutants in urban environments. The aim of this article is to show that forecast of daily airborne pollution using support vector machines (SVM) is feasible in regression mode. Results are presented using data measurements of Particulate Matter of aerodynamical size on the order of 10 and 2.5 micrograms (PMx) in London-Bloomsbury at south England.**

Index Terms— **Particulate matter, Support Vector Machines, PMx, airborne pollution, forecast.**

I. Introduction

In recent times, urban air pollution has been a growing problem especially for urban communities. Size, shape and chemical properties govern the lifetime of particles in the atmosphere and the site of deposition within the respiratory tract. Health effects differ upon the size of airborne particulates. In this contribution, PM10 (particles less or equal than 10 micrometers) and PM2.5 (particles less or equal than 2.5 micrometers) are considered due to its effect on human health, according to several authors [1-6] This is the primary reason this research has been done; to monitor, and model the levels and spread of PMx in urban environments. In previous contributions, it has been shown that forecast of concentration levels of PM10 may be possible by using other techniques such as neural networks and various fuzzy clustering algorithms [7-8]. However, even though these works have shown that is feasible to robust model the non-linear behavior of the system, a more robust model is needed with an enhanced method to reduce the error between the raw data and the model. For this reason, support vector machines (SVM) are chosen for this work. In this appraisal, the modeling will be carried out using support vector machines working in regression mode. Support vector machines are a recent statistical learning technique, based on machine learning and generalization theories, it implies an idea and could be considered as a method to minimize the risk [9]. Also, a generalization capability makes possible their application to modeling dynamical and non-linear data sets.

II. Support Vector Machines

A. Theory of support Vector machines for regression

The support vector machines (SVM) theory, was developed by Vapnik in 1995, and is applied in many machine-learning applications such as object classification, time series prediction, regression analysis and pattern recognition. Support vector machines (SVM) are based on the principle of structured risk minimization (SRM) [9-10].

In the analysis using SVM, the main idea is to map the original data $\boldsymbol{x}$ into a feature space $\boldsymbol{F}$ with higher dimensionality via non-linear mapping function ϕ, which is generally unknown, and then carry on linear regression in the feature space [8]. Thus, the regression approximation addresses a problem of estimating function based on a given data set (where x_i represent the input vectors, d_i are the desired values), which is produced from the ϕ function. SVM method approximates the function by:

$$y = \sum_{i=1}^{m} w_i \phi_i(x) + b = w\phi(x) + b \tag{1}$$

where $w = [w_1, \dots, w_m]$ represent the weights vector, b are the bias coefficients and $\phi(x) = [\phi_1(x), \dots, \phi_m(x)]$ the basis function vector.

The learning task is transformed to the weights of the network at minimum. The error function is defined through the ε-insensitive loss function, $L_\varepsilon(d, y(x))$ and is given by:

$$L_\varepsilon(d, y(x)) = \begin{cases} |d - y(x)| - \varepsilon & |d - y(x)| \geq \varepsilon \\ 0 & others \end{cases} \tag{2}$$

The solution of the so defined optimization problem is solved by the introduction of the Lagrange multipliers α_i, α_i^* (where i=1,2,…,k) responsible for the functional constraints defined in Eq. 2. The minimization of the Lagrange function has been changed to the dual problem [9]:

$$\phi(\alpha, \alpha^*) = \left[\sum_{i=1}^{k} d_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{k} (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} (\alpha_i, \alpha_i^*)(\alpha_j, \alpha_j^*) K(x_i, x_j) \right] \tag{3}$$

ISBN: 978-607-95060-3-2
Supported by the IEEE Aguascalientes Section and the Instituto Tecnologico de Aguascalientes

With the constraints

$$\sum_{i=1}^{k}(\alpha_i,\alpha_i^*)=0, \qquad 0\le\alpha_i\le C, 0\le\alpha_i^*\le C \tag{4}$$

Where C is a regularized constant that determines the trade-off between the training risk and the model uniformity.
According to the nature of quadratic programming, only those data corresponding to non-zero $(\alpha_i-\alpha_i^*)$ pairs can be referred to support vectors (Nsv). In Eq. 3 $K(x_i , x_j)=\phi(x_i)*\phi(x_j)$ is the inner product kernel which satisfy Mercer's condition [11] that is required for the generation of kernel functions given by:

$$K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle \tag{5}$$

Thus the support vectors associates with the desired outputs $y(x)$ and with the input training data x can be defined by:

$$y(x)=\sum_{i=1}^{Nsv}(\alpha_i,\alpha_i^*)K(x,x_i)+b \tag{6}$$

Where x_i are learning vectors. Leading us to a SVM architecture and are also founded in [7][8][12].

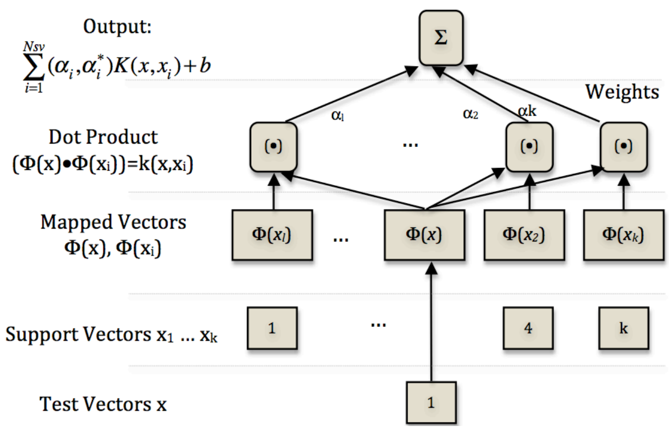

Fig. 1. Support Vector Machine Architecture.

B. Kernel function

The use of an appropriate kernel is the key feature in support vector applications, since it provide the capability of mapping non-linear data into "feature" spaces that in essence are linear, then an optimization process can be applied as in the linear case.

1) The Gaussian Kernel

The Gaussian kernel function is defined in [9-11] Eq. 7.

$$K(x_i,x_j)=\exp\left(\frac{-\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{7}$$

The Gaussian kernel process delivers an estimate for the reliability of the prediction in the form of the variance of the predictive distribution and the analysis can be used to estimate the evidence in favor of a particular choice of covariance function. The covariance or kernel function can be seen as a model of the data, thus providing a principled method for model selection [12-13].

C. The free parameters

Other important issues in support vector applications are the selection of free parameters such as the coefficient of C, the value of error ε it determine the margin within which error is neglected and in the Gaussian kernel function the value of variances σ[13-15].

D. The quadratic programming problem

The SVM training works flawlessly for not too large data sets. However, when the number of data points is large, over 2,000, the Quadratic Programming (QP), problem becomes extremely difficult to solve with standard QP solvers and methods [11-16]. In the study case of this survey, the number of data points is 365, where each data point represents the daily average of PMx concentration. Therefore the analysis and solving of the QP problem is not considered in the scope of this survey. See Fig 2.

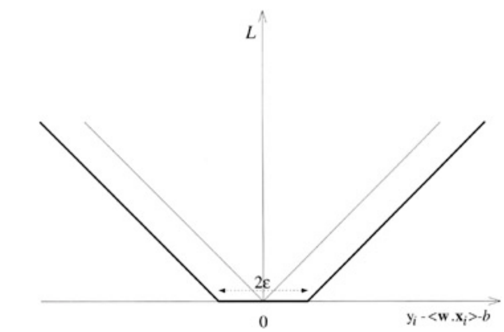

Fig 2a. The linear ε-insensitive loss for zero and non-zero ε

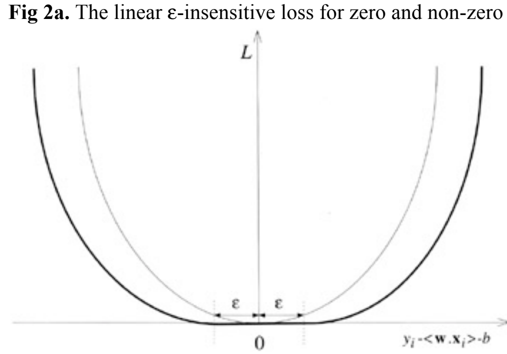

Fig 2b. The quadratic ε-insensitive loss for zero and non-zero ε

Fig 2. The ε-insensitive loss functions for zero and non-zero ε

III. Methodology

The proposed Methodology have been taken from [3-4], such this works provides the general steps to make pollutants modeling and predictions by using SVM working in regression mode. In this survey only a Gaussian kernel functions is used and implemented the main reasons to choose a Gaussian kernel functions are, it has been widely used in the literature [3][4][13][14][15], a Gaussian distribution provides a natural representation of the system behavior [13][15]. The aim of this survey is to show the relations between kernel Gaussian parameters and the obtained SVM models.

ISBN: 978-607-95060-3-2
Supported by the IEEE Aguascalientes Section and the Instituto Tecnologico de Aguascalientes

In order to perform an appropriate design, train, and testing of SVM this article describes a generic methodology based in a review of [3-4].

(a) Preprocess the input data and select the most relevant features, scale the data in the range [−1, 1], and check for possible outliers.
(b) Select an appropriate kernel function that determines the hypothesis space of the decision and regression function.
(c) Select the parameters of the kernel function the variances of the Gaussian kernels.
(d) Choose the penalty factor C and the desired accuracy by defining the ε-insensitive loss function.
(e) Validate the model obtained on some previously, during the training, unseen test data, and if not pleased iterate between steps (c) (or, eventually b) and (e).

IV. Discussion

The fundamental reason for considering SVM working in regression mode as an approach for PMx modeling is the non-linear aspect of the application.

There is no predetermined heuristic for the choice of free parameters and design for the SVM, many applications appear to be specific, in order to improve the SVM performance trough the automatic adjustment of free parameters.

Using SVM on real time applications appear to be rather complex since of the computational demands of the deriving results.

V. Experimental Results

The Support Vector method can be applied to the case of regression, maintaining all the main features that characterize the system behavior. An SVM in a kernel-induced feature space learns a non-linear function while the capability of the system is controlled by a set of parameters that does not depend on the dimensionality of the space. In this section presents a set of results and simulations by using the proposed regression SVM model approach with Gaussian kernel functions and standard nonlinear data sets of PMx. During 2009, simulations were carried out using the proposed SVM model. The σ values were modified to 1 and 2. Likewise, the ε values were modified to 7,11 and 13. For every case study, the normalized value C remained content to a value of 100. Also is observed that the error rate of standard SVM varies wildly depending on different values of SVM free-parameters and kernel parameters.

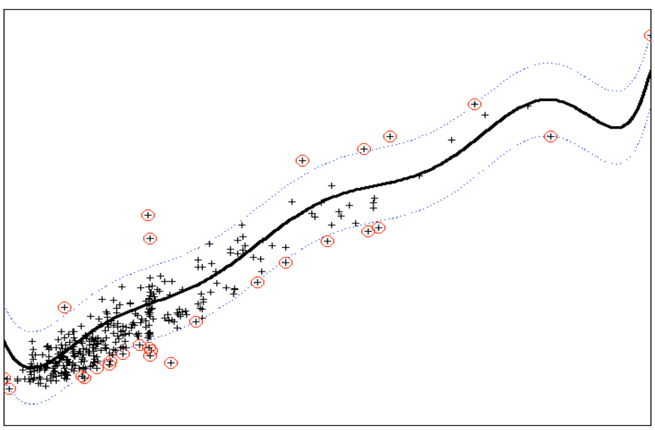

Fig 3.a. SVM estimated with free parameters of ε = 7, σ = 1.

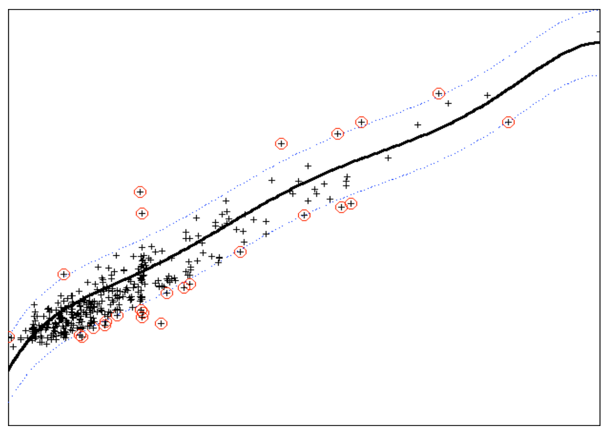

Fig 3.b. SVM estimated with free parameters of ε = 7, σ = 2.

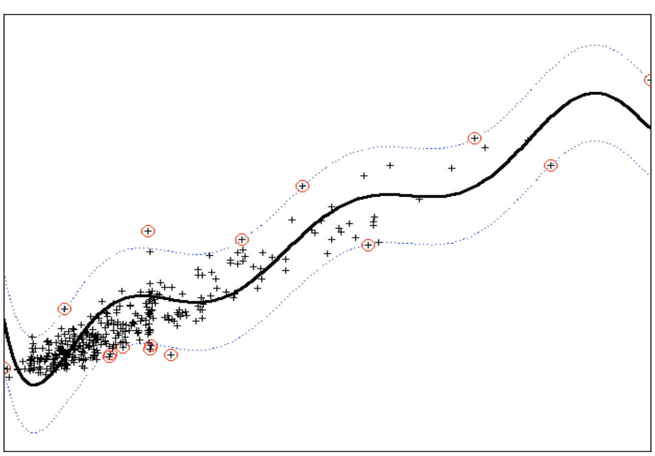

Fig 3.c. SVM estimated with free parameters of ε = 11, σ = 1.

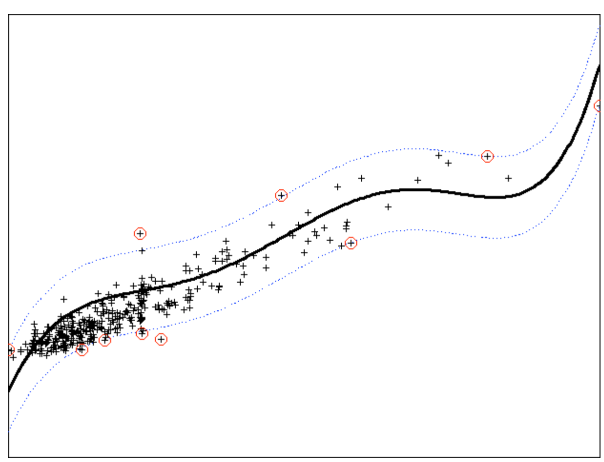

Fig 3.d. SVM estimated with free parameters of ε = 11, σ = 2.

ISBN: 978-607-95060-3-2
Supported by the IEEE Aguascalientes Section and the Instituto Tecnologico de Aguascalientes

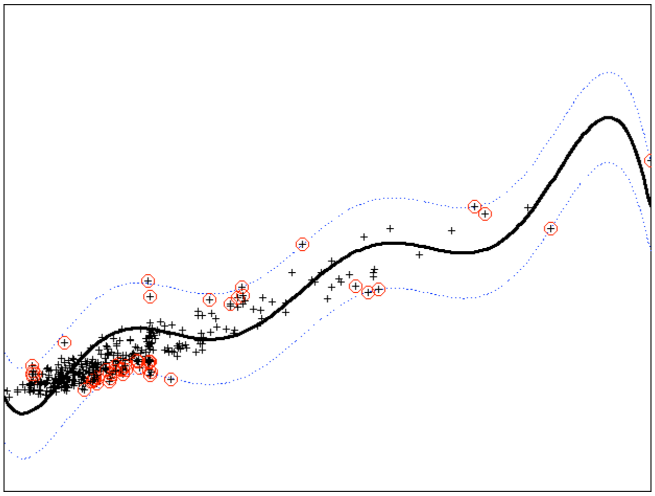

Fig 3.e. SVM estimated with free parameters of ε = 13, σ = 1.

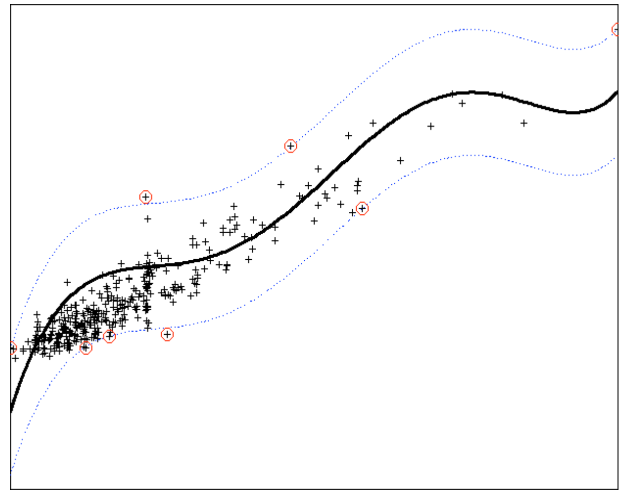

Fig 3.f. SVM estimated with free parameters of ε = 13, σ = 2.

Fig. 3. Summary of model estimations using SVM.

Figure 3 shows a summary of the results with the Support vector machine (in red circles), the raw data (black cross) and the behavior of the data (solid black line). These results show that the best results are obtained with σ of 2 and an ε of 13 (figure 2f) due to the small number of SVMs and small error rate, whilst the worst-case scenario is obtained with a σ of 1 and an ε of 13 (figure 3e), since a large number of 295 SVMs is obtained.
From these results, it can be concluded that for this case study a σ of 1 gives a similar number of SVMs with respects to the number of data points. This exponentially increases the computational cost, making it unfeasible to calculate it.

VI. CONCLUSION

This survey has presented a modeling method of the daily atmospheric pollution by applying the support vector machine with Gaussian kernel functions working in regression mode. The application of SVM has enabled to obtain a good accuracy in modeling pollutant concentration of both PM10 and PM2.5.
The methods, techniques and alternatives offered in the SVM field provides a flexible and scalable tool for implementing sophisticated solutions with implied dynamical and non-linear data. It is noteworthy to point that the SVM guarantees this global minimum solution and a good feature of generalization.
Furthermore, implementing other kernel functions like polynomial, wavelet and hybrid functions may be implemented for future contributions.

ACKNOWLEDGMENT

The authors would like to thank the Air Quality Archive hosted by AEA Energy & Environment, on behalf of the United Kingdom Department for Environment, Food & Rural Affairs and the Devolved Administrations (DEFRA). Also, the authors would like to acknowledge the financial support of the Mexican government via SEP-PROMEP /103-5/09/4100 Project.

REFERENCES

[1] Lall R, Kendall M, Ito K, Thurston G. D.: Estimation of Historical Annual PM2.5 Exposures for Health Effects Assessment, Atmospheric Environment, Vol. 38, 2004 pp. 5217--5226.
[2] Querol X., Alastvey A., Ruiz C. R., et al.: Speciation and origin of PM10 and PM 2.5 in selected European cities, Atmospheric Environment, Vol. 38, 2004 pp. 6547--6555.
[3] F. Wang, D.S. Chen, S.Y. Cheng, J.B. Li, M.J. Li, Z.H. Ren, "Identification of regional atmospheric PM10 transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis", Environmental Modelling & Software 25 2010 , pp. 927-934.
[4] Ling – Yan He, Minhu ,Yuan – Hangzhang, Xiao - Fenghuang, Ting – Tin Gyao, "Fine Particle Emissions from On-Road Vehicles in the Zhujiang Tunnel, China", Environ. Sci. Technol. 2008, 42, pp. 4461–4466.
[5] David J. Briggs, Kees de Hoogh, Chloe Morris, John Gulliver, "Effects of travel mode on exposures to particulate air pollution", Environment International, 34, 2008, pp. 12–22.
[6] Giorgio Zamboni, Massimo Capobianco, Enrico Daminelli, "Estimation of road vehicle exhaust emissions from 1992 to 2010 and comparison with air quality measurements in Genoa, Italy", Atmospheric Environment 43, 2009, pp. 1086–1092.
[7] Collazo-Cuevas J.I., Aceves-Fernandez M.A., Gorrostieta-Hurtado E., Pedraza-Ortega J.C., Sotomayor-Olmedo A.2, Delgado-Rosas M., "Comparison between Fuzzy C-means Clustering and Fuzzy Clustering Subtractive in urban air Pollution", CONIELECOMP 2010, 20th International Conference on Electrical Communications, pp. 174-179.
[8] M. A. Aceves-Fernandez, J. I. Collazo-Cuevas, E. Gorrostieta, C. Pedraza-Ortega, J. M. Ramos, S. Canchola. Prognosis of Urban Air Pollution by using a Fuzzy Clustering System In Northwest England journal in Computing Science, ISSN 1870-4069.
[9] Vapnik, V.: The Nature of Statical Learning Theory. Springer-Verlang, New York. 1995.
[10] Vapnik, V., Golowich, S., Smola A.: Support method for function approximation regression estimation, and signal processing. Advance in Neural Information Processing System 9. MIT Press, Cambridge, MA. 1997
[11] Schölkfopf B.: Smola A. J.: and Burges C.: Advances in Kernel Methods –Support Vector Learning. Cambridge, M.A.: MIT Press. 1999
[12] Osuna, E., R. Freund, F. Girosi.: Support vector machines: Training and applications. AI Memo 1602, Massachusetts Institute of Technology, Cambridge, MA 44. (1997
[13] Cristianini, N., Shawe-Taylor, J., An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, UK 2000.

[14] S. Osowski and K. Garanty, "Forecasting of the daily meteorological pollution using wavelets and support vector machine," Engineering Applications of Artificial Intelligence, vol. 20, no. 6, pp. 745-755, September 2007. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2006.10.008.

[15] I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," Computational Intelligence Magazine, IEEE, vol. 4, no. 2, pp. 24-38, 2009. [Online]. Available: http://dx.doi.org/10.1109/MCI.2009.932254

[16] W. Lu and W. Wang, "Potential assessment of the support vector machine method in forecasting ambient air pollutant trends," Chemosphere, vol. 59, no. 5, pp. 693-701, April 2005. [Online]. Available: http://dx.doi.org/10.1016/j.chemosphere.2004.10.032